ТЕОРИЯ ВЕРОЯТНОСТЕЙ
И
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Лекция 10

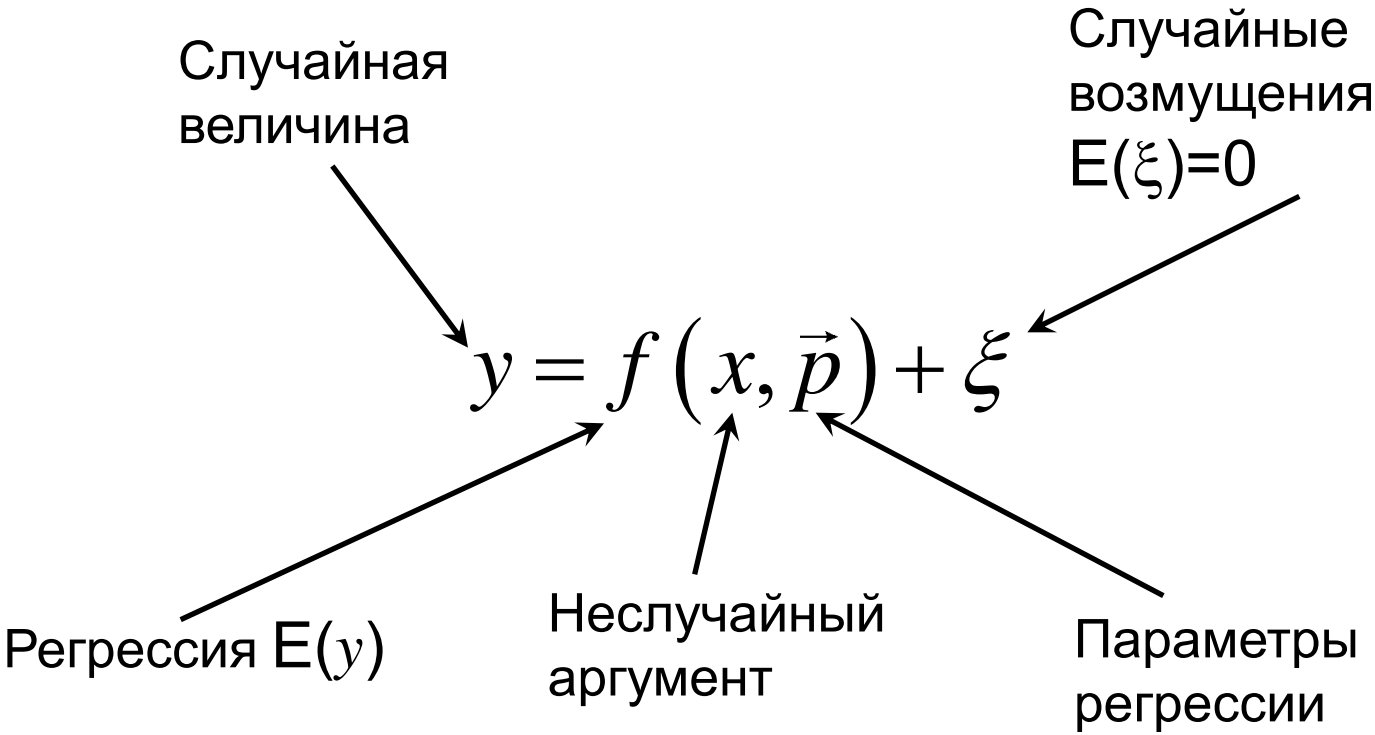
Изначальное содержание понятия регрессии:
Регре́ссия (лат. *regressio* — обратное движение, отход)

Общее понимание регрессии в математической статистике:

Пусть имеется пара случайных переменных (X, Y).
Регрессия – это функция $g\left(x\right) = E\left(Y \mid X = x\right)$

то есть, условное математическое ожидание случайной переменной Y при условии, что другая случайная переменная X приняла значение x.

Далее регрессией будем называть зависимость математического ожидания случайной величины от неслучайных переменных.



Как правило, вид функции f задаётся исследователем.
Задача: найти значения вектора $\vec{p}$, наилучшим образом соответствующие данным, полученным в эксперименте, когда значения x задаются, а значения y измеряются.

Если f - полином, то

$$f\left(x,\vec{p}\right)=a_0+a_1x+a_2x^2+a_3x^3+\ldots+a_nx^n$$

где n - конечное и небольшое.

Вектор параметров:

$$\vec{p}=\left(a_0,a_1,a_2,a_3,\ldots,a_n\right)$$

Это пример линейной регрессии.

Общий случай линейной регрессии:

$$f\left(x,\vec{p}\right)=\sum_{l=1}^{L}p_l f_l\left(x\right)$$

f_l - базисные функции, не зависящие от p_l

В случае полинома

$$f_1\left(x\right)=x^0=1,\quad f_2\left(x\right)=x^1=x,\quad f_3\left(x\right)=x^2\quad \ldots\ f_L\left(x\right)=x^{L-1}$$

Вектор параметров $\vec{p}$ должен быть таким, чтобы кривая регрессии наилучшим образом прошла между экспериментальными точками y

Принцип максимального правдоподобия

Предполагается, что $$w(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\xi^2}{2\sigma^2}}$$

Отсюда $$w(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left[y - f(x,\vec{p})\right]^2}{2\sigma^2}}$$

(x_n, y_n) – пары экспериментальных данных

Вероятность попадания y_n в интервал $y_n - \frac{1}{2}\Delta y < y < y_n + \frac{1}{2}\Delta y$:

$$w(y_n) = \frac{\Delta y}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left[y_n - f(x_n, \vec{p})\right]^2}{2\sigma^2}}$$

Для N экспериментальных точек:

$$w(y_1, y_2, ..., y_N) = \prod_{n=1}^{N} w(y_n) = \left(\frac{\Delta y}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{n=1}^{N} e^{-\frac{\left[y_n - f(x_n, \vec{p})\right]^2}{2\sigma^2}}$$

Функция правдоподобия:

$$LH = \prod_{n=1}^{N} e^{-\frac{\left[y_n - f\left(x_n, \vec{p}\right)\right]^2}{2\sigma^2}} = \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[y_n - f\left(x_n, \vec{p}\right)\right]^2\right\}$$

Функция правдоподобия максимальна при

$$S = \sum_{n=1}^{N}\left[y_n - f\left(x_n, \vec{p}\right)\right]^2 = \min$$

На этом основан метод наименьших квадратов:
найти значения параметров $\vec{p}$,
при которых S минимальна

Метод наименьших квадратов:

$$\frac{\partial}{\partial p_l} S = -2\sum_{n=1}^{N}\left\{\left[y_n - f\left(x_n, \vec{p}\right)\right]\frac{\partial}{\partial p_l} f\left(x_n, \vec{p}\right)\right\} = 0, \quad l = 1,2,...,L$$

Общий случай
линейной регрессии:

$$f\left(x, \vec{p}\right) = \sum_{l=1}^{L} p_l f_l\left(x\right)$$

Тогда

$$\frac{\partial}{\partial p_l} f\left(x, \vec{p}\right) = f_l\left(x\right)$$

Подставляем:

$$\sum_{n=1}^{N}\left\{\left[y_n-\sum_{m=1}^{L}p_m f_m\left(x_n\right)\right]f_l\left(x_n\right)\right\}=0,\ \ l=1,2,...,L$$

$$\sum_{n=1}^{N}y_n f_l\left(x_n\right)-\sum_{n=1}^{N}\left\{f_l\left(x_n\right)\sum_{m=1}^{L}p_m f_m\left(x_n\right)\right\}=0,\ \ l=1,2,...,L$$

$$\sum_{m=1}^{L}p_m\left\{\sum_{n=1}^{N}f_l\left(x_n\right)f_m\left(x_n\right)\right\}=\sum_{n=1}^{N}y_n f_l\left(x_n\right),\ \ l=1,2,...,L$$

Обозначим:

$$\sum_{n=1}^{N} f_l\left(x_n\right) f_m\left(x_n\right) = a_{lm}, \quad \sum_{n=1}^{N} y_n f_l\left(x_n\right) = b_l$$

Система линейных уравнений
для параметров линейной регрессии:

$$\sum_{m=1}^{L} a_{lm} p_m = b_l, \quad l = 1,2,...,L$$

Векторно-матричная форма:

$$A\vec{p} = \vec{b}$$

Решение:

$$\vec{p} = A^{-1}\vec{b}$$

Расчет матрицы ошибок значений параметров:

$$RSS = S_{\min} \qquad s_y^2 = \frac{RSS}{N - L}$$

Матрица ошибок:

$$D_{\vec{p}} = s_y^2 A^{-1}$$

$$D_{\vec{p}} = \begin{pmatrix} \sigma_{p_1}^2 & \operatorname{cov}(p_1, p_2) & \dots & \operatorname{cov}(p_1, p_L) \\ \operatorname{cov}(p_2, p_1) & \sigma_{p_2}^2 & \dots & \operatorname{cov}(p_2, p_L) \\ \dots & \dots & \dots & \dots \\ \operatorname{cov}(p_L, p_1) & \operatorname{cov}(p_L, p_2) & \dots & \sigma_{p_L}^2 \end{pmatrix}$$

$$r(p_i, p_j) = \frac{\operatorname{cov}(p_i, p_j)}{\sigma_{p_i} \sigma_{p_j}}$$

Алгоритм вычисления параметров линейной регрессии

1) Выбор базисных функций

Затем вычисление следующих величин

2) Матрица A и вектор правый частей $\vec{b}$

3) Матрица A^{-1}

4) Параметры регрессии $\vec{p}$

5) Остаточная сумма квадратов отклонений RSS и величина s_y^2

6) Матрица ошибок $D_{\vec{p}}$